An algorithm to detect Copy Number Aberrations in cancer genomes of tumour specimens.

Arief Gusnanto, Stefano Berri, Henry M. Wood and Pamela Rabbitts
The cancer genome is often aneuploid.

Hartwell and Kastan. Science, 1994
Detecting abnormalities

Why?
- Molecular characterisation and classification of tumours
- Diagnostic, prognostic and predictive tool
- Understand the biology of cancer

How?
- CGH
- aCGH (BAC or oligo)
- SNP microarray
- “NextGen” Sequencing
  - Tuneable resolution/cost
  - Re-use of data
  - Flexible platform
  - Technical independence Test – Control
  - Might become very cheap
Copy number by “NextGen” Sequencing
“NextGen” sequencing and reads mapping.
Counting number of sequences for each window

Tumour: 2 3 3 2 2 3 2

Normal: 2 2 2 2 2 3 2
Counting number of sequences for each window

<table>
<thead>
<tr>
<th>Genomic location</th>
<th>Ratio (Copy num)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumour</td>
<td>2 3 3 2 2 3 2</td>
</tr>
<tr>
<td>Normal</td>
<td>2 2 2 2 2 3 2</td>
</tr>
</tbody>
</table>

Ratio values are as follows:
- 2 (4)
- 1.5 (3)
- 1 (2)
- 0.5 (1)
- 0 (0)
Toward the real data

Distribution of read counts. Simulated Data, 3M reads
Copy number from simulated Data
Copy number from simulated Data

SmoothSeg. Huang et al, 2007
Resolution and noise go together!
Different number of total reads

Reference genome

Tumour

Normal
Total number of reads varies.

Copy number from simulated Data.
Unequal number of total reads
Normalization. A crucial step

Test before normalization

Control
Normalization. A crucial step

Test before normalization

Test after median normalization

Control

Control
Normalization. A crucial step

Copy number from simulated data after median normalization.
The cancer genome is often aneuploid

Amplifications
Deletions
Normal

Many amplifications and deletions!

Hartwell and Kastan. Science, 1994
Patient’s tumour samples

- Contamination with stroma, inflammatory cells...
The real samples. A lot noisier

- Some sequences cannot be aligned (repeated regions)
- GC content bias
- Unequal number of total reads.
- Extra noise of unknown origin
The median might be meaningless

- Asymmetric distribution
- “Flat top”
- long tail
Median normalisation

Median-normalized data. Patient's specimens
Trade resolution for noise

Median-normalized data. Patient's specimens

Copy Number, Window 500 Kbp wide

Chromosome 3, Mbp

0 50 100 150 200

0.0 1.0 2.0 3.0
Discrete data normalisation

Patient's sample, segmented data
Discrete data normalisation

Patient's sample, segmented data

Density

Ratio Test/Normal

0 (0) 0.5 (1) 1 (2) 1.5 (3) 2 (4) 6
Discrete data normalisation

Patient's sample, segmented data
Patient's sample, segmented data

\[
\text{Contamination} = \frac{\text{Tumour tissue}}{\text{Total tissue}} = 51\%
\]
Discrete normalization. Patient's specimens

Copy number. Window 50 Kbp wide

Chromosome 3, Mbp
Conclusions

• Develop a novel normalisation method for “NextGen” data that can cope with
  ✓ Highly abnormal genomes
  ✓ Tumour samples contaminated by normal cells

• We can estimate contamination percentage.
Limits

• Contamination is allowed, but otherwise the tumour should be homogeneous.

• Process might require human supervision when calling discrete states.
Acknowledgements

Department of Statistics, School of Mathematics
Arief Gusnanto

Leeds Institute of Molecular Medicine, St. James’s University Hospital
Pamela Rabbitts, Henry M. Wood
Catherine Daly, Ornella Belvedere
Joanne Morgan, Graham Taylor

Yorkshire Cancer Research